

# Sparse Orthonormalized Partial Least Squares

Marcel van Gerven

Tom Heskes

*Radboud University Nijmegen, ICIS, P.O.Box 9010 6500GL Nijmegen*

## Abstract

Orthonormalized partial least squares (OPLS) is often used to find a low-rank mapping between inputs  $X$  and outputs  $Y$  by estimating loading matrices  $A$  and  $B$ . In this paper, we introduce sparse orthonormalized PLS as an extension of conventional PLS that finds sparse estimates of  $A$  through the use of the elastic net algorithm. We apply sparse OPLS to the reconstruction of presented images from BOLD response in primary visual cortex. Sparse OPLS finds solutions with low reconstruction error which are easy to interpret due to the sparseness of the loading matrix  $B$ . Moreover, the elastic net algorithm is generalized to allow for coupling constraints that induce a spatial regularization.

## 1 Introduction

Suppose we are given a dataset with high-dimensional inputs  $X$  and outputs  $Y$  and somehow would like to learn a sparse low-rank mapping between the two. Sparse, because we expect not too many inputs to be relevant, and low-rank, because we expect there to be some lower-dimensional features that (given the limited amount of data we have) explains the output. As a motivating example, consider neuroimaging data obtained when subjects are presented different images. A sparse low-rank mapping could be used to identify a low-dimensional representation which allows presented images to be reconstructed from measured neuroimaging data. This task is an example of ‘brain-reading’ using multivariate methods which received a lot of recent attention from the neuroscience community (e.g., [8, 9, 10, 6]).

There are many different approaches to solve the low-rank mapping problem. Canonical correlation analysis (CCA) tries to find projections of  $X$  and  $Y$  onto lower-dimensional subspaces with maximum correlation [7]. CCA can be kernelized [1], has a probabilistic interpretation [16] (similar to probabilistic interpretations of PCA [15]), and can be sparsified [5]. CCA has been used successfully in the context of brain-reading [4]. However, a disadvantage of CCA is that it treats  $X$  and  $Y$  symmetrically, i.e., it is not optimized towards predicting  $Y$  given  $X$ , which in many cases appears to be the most natural paradigm. Partial least squares (PLS) does something similar to CCA, but then aims to maximize covariance instead of correlation. There are many variants of PLS, depending on how data is deflated when a new projection is found [11, 13].

In this paper, we introduce a novel approach to partial least squares that makes use of the equivalence between partial least squares and heteroencoders [12]. By adding both an L1 penalty term and a (Laplacian) ridge regression term, we enforce sparsity and smoothness between neighbouring voxels. Our line of reasoning largely follows that of Zou et al. [18], who suggested a sparsified variant of PCA through its relation with an autoencoder. An alternative, more indirect approach towards sparse partial least squares, is taken in the recent paper of Chun and Keleş [2] where sparsity is introduced by imposing an L1-penalty onto a surrogate of the (unsparsified) partial least squares solutions. We demonstrate the usefulness of our approach on a brain-reading dataset where the task was to reconstruct presented images from measured BOLD response.

## 2 Partial least squares

We start by discussing how PLS is related to minimizing a sum-squared error using a heteroencoder network [12]. A heteroencoder model takes input  $\mathbf{x}$  and outputs  $\mathbf{y}$  through

$$\begin{aligned}\mathbf{z} &= B^T \mathbf{x} \\ \mathbf{y} &= A \mathbf{z},\end{aligned}$$

where  $A$  and  $B$  are  $M \times K$  and  $P \times K$  loading matrices, respectively. Typically, the number of hidden units  $K$  is much smaller than the number of inputs  $P$  or outputs  $M$ , such that we have a bottleneck network. Note that this model is unidentifiable: a model with  $\tilde{A} = AC^{-1}$  and  $\tilde{B} = BC^T$  is completely equivalent for any invertible  $K \times K$  matrix  $C$ . To partly get rid of this unidentifiability, we can use the orthonormality constraint  $A^T A = I_K$ . The model is then identifiable up to row-wise sign flips and permutations.

Given a dataset with input-output pairs  $(\mathbf{x}_n, \mathbf{y}_n)$ , we would like to find those  $A$  and  $B$  that minimize the sum-squared error plus a ridge penalty on each of the columns  $B_k$  of  $B$  under the constraint  $A^T A = I_K$ . I.e., we aim to find the minimizer  $(\hat{A}, \hat{B})$  of

$$\frac{1}{2N} \sum_n \|\mathbf{y}_n - AB^T \mathbf{x}_n\|_2^2 + \frac{1}{2} \lambda \sum_k B_k^T B_k \quad (1)$$

subject to  $A^T A = I_K$ . Fixing  $A$  and solving for  $B$ , we obtain

$$\hat{B}(A) = [\Sigma_{xx} + \lambda I_P]^{-1} \Sigma_{xy} A$$

with sample covariance matrices  $\Sigma_{xx} \equiv \frac{1}{N} \sum_n \mathbf{x}_n \mathbf{x}_n^T$  and  $\Sigma_{xy} \equiv \frac{1}{N} \sum_n \mathbf{x}_n \mathbf{y}_n^T$ . After substitution of this solution back into (1) and some simplifications, we arrive at

$$\hat{A} = \arg \max_A \text{Tr} \left( A^T \Sigma_{yx} [\Sigma_{xx} + \lambda I_P]^{-1} \Sigma_{xy} A \right)$$

subject to  $A^T A = I_K$ . This is a standard eigenvalue problem, which implies that  $\hat{A}$  is made up of the eigenvectors of  $\Sigma_{yx} [\Sigma_{xx} + \lambda I_P]^{-1} \Sigma_{xy}$  corresponding to the  $K$  largest eigenvalues.

In the limit  $\lambda \rightarrow 0$ , the input covariance matrix dominates the diagonal matrix and we arrive at the same solution as orthonormalized PLS (OPLS). In the limit  $\lambda \rightarrow \infty$ , on the other hand, we obtain the criterion commonly attributed to standard PLS, i.e., maximize the covariance between the input and the output.

## 3 Sparse OPLS

Zou et al. [18] follow a similar approach by exploiting the equivalence between principal component analysis and autoencoders, which corresponds to setting  $Y = X$  in the above. On top of the quadratic penalty term they add an additional L1 penalty term on the elements of  $B$  which gives sparse solutions. Inspired by their work, we propose to do the same, but then for the hetero-encoder instead of for the auto-encoder, yielding a sparse OPLS algorithm. More specifically, we consider the following optimization criterion

$$\frac{1}{2N} \sum_n \|\mathbf{y}_n - AB^T \mathbf{x}_n\|_2^2 + \nu \sum_k \|B_k\|_1 + \frac{1}{2} \sum_k B_k^T \Lambda B_k \quad (2)$$

where, compared with (1), we added an L1 penalty and generalized the ridge penalty to allow for a matrix  $\Lambda$  instead of just a constant. Following the same line of reasoning as in [18], we arrive at the following procedure for updating  $A$  and  $B$ .

- Fix  $A$ , reconstruct  $\mathbf{z}_n = A^T \mathbf{y}_n$  for all  $n$  and solve

$$\hat{B} = \arg \min_B \frac{1}{2N} \sum_n \|\mathbf{z}_n - B^T \mathbf{x}_n\|_2^2 + \nu \sum_k \|B_k\|_1 + \frac{1}{2} \sum_k B_k^T \Lambda B_k. \quad (3)$$

- Fix  $B$ , reconstruct  $\mathbf{z}_n = B^T \mathbf{x}_n$  for all  $n$  and solve

$$\hat{A} = \arg \min_A \frac{1}{2N} \sum_n \|\mathbf{y}_n - A \mathbf{z}_n\|_2^2$$

subject to  $A^T A = I_K$ . In terms of the singular value decomposition  $\Sigma_{yz} = UDV^T$ , the solution is simply  $\hat{A} = UV^T$  or, equivalently,  $\hat{A} = \Sigma_{yz} (\Sigma_{yz}^T \Sigma_{yz})^{-1/2}$ .

In practice, it makes sense to build the model sequentially, starting from a model with a single latent variable and adding new latent variables one by one. In that case, the optimization problem (3) boils down to solving a sequence of so-called elastic nets [17] for each  $B_k$  separately.

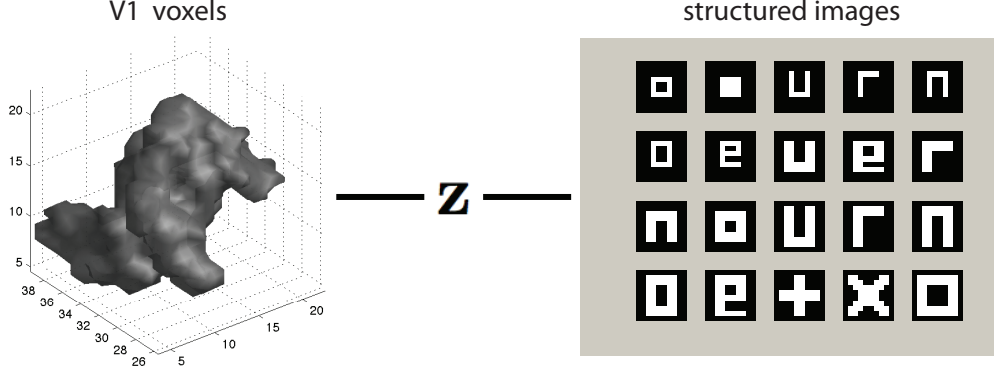


Figure 1: A sparse low-rank mapping between V1 BOLD response and presented images is captured through the latent variables  $\mathbf{z}$  of the sparse OPLS model.

Let us consider the special case in which the number of inputs  $P$  is larger than the number of samples  $N$ . In that case, no matter what matrix  $A$  is chosen, with  $\nu = 0$  and  $\Lambda = \mathbf{0}$ , it is always perfectly possible to fit the reconstructed  $Z$ . The choice of  $A$  is then solely dominated by the characteristics of  $Y$  and it is easy to show that the optimal  $A$  consists of the principal eigenvectors of  $\Sigma_{yy}$ . When  $\nu \neq 0$  and/or  $\Lambda \neq \mathbf{0}$ , the optimal solution will deviate somewhat, but it is still to be expected that  $A$  is close to the principal eigenvectors of  $\Sigma_{yy}$ . It therefore makes sense to initialize  $A$  to those eigenvectors.

In order to solve the elastic net, we follow a strategy which generalizes the approach of [3]. For each  $B_k$ , denoted by  $\beta$  in the following, we consider minimizing a function

$$R(\beta) = \frac{1}{2N} \sum_n (z_n - \beta^T \mathbf{x}_n)^2 + \nu \|\beta\|_1 + \frac{1}{2} \beta^T \Lambda \beta \quad (4)$$

with respect to  $\beta$  through coordinate descent. Fixing  $\beta_j = \tilde{\beta}_j$  for  $j \neq i$  and differentiating with respect to  $\beta_i$ , we obtain

$$\frac{\partial R}{\partial \beta_i} = -T_i + \nu \text{sign}(\beta_i) + [S_{ii} + \Lambda_{ii}] \beta_i + \sum_{j \neq i} [S_{ij} + \Lambda_{ij}] \tilde{\beta}_j$$

with  $T_i \equiv \frac{1}{N} \sum_n z_n x_{in}$  and  $S_{ij} \equiv \frac{1}{N} \sum_n x_{in} x_{jn}$ . If we further define  $V_i \equiv T_i - \sum_{j \neq i} Q_{ij} \tilde{\beta}_j$  and  $U_i \equiv S_{ii} + \Lambda_{ii}$  with  $Q_{ij} \equiv S_{ij} + \Lambda_{ij}$ , we obtain the solution

$$\beta_i = \begin{cases} 0 & \text{if } |V_i| \leq \nu \\ \frac{V_i - \nu}{U_i} & \text{if } V_i > \nu \\ \frac{V_i + \nu}{U_i} & \text{if } V_i < -\nu \end{cases}$$

So, a simple procedure would be the following:

- Start from  $\beta = \mathbf{0}$  and  $V_i = T_i$  for all  $i$ .
- Iterate by picking a variable  $i$  and checking whether  $\beta_i$  changes (typically when  $|V_i| \geq \nu$ , but also when  $\beta_i$  changes from nonzero back to zero). If it does, we not only update  $\beta_i$ , but also  $V_j$  for all  $j \neq i$ .

- Repeat until convergence.

Note that we only need elements  $Q_{i\bullet}$  (and thus  $S_{i\bullet}$ ) when variable  $i$  at some point enters the equation. In particular in problems with many input variables, of which only some are expected to enter, it pays not to compute  $Q_{ij}$  beforehand, but only (once) on the fly when it is needed and then store it for later usage.

Finally, note that Eq. (4) represents the elastic net only when  $\Lambda = \lambda I_P$ . In general, a non-diagonal  $\Lambda$  allows coupling constraints to be formulated between variables. This idea is similar in spirit to that of the fused lasso [14] although, in that case, an additional  $L_1$  instead of  $L_2$  regularizer is used to realize the coupling. In our experiments, we use a non-diagonal  $\Lambda$  in order to induce a spatial regularization of the estimated regression coefficients.

## 4 Brain reading with sparse OPLS

### 4.1 Experimental data

We apply the sparse OPLS algorithm to the reconstruction of presented images from measured BOLD response. This data has been described in [9] and can be downloaded from [http://www.cns.atr.jp/~yoichi\\_m](http://www.cns.atr.jp/~yoichi_m). The dataset consists of measured BOLD response in occipital cortex for 440 random images and 480 structured images (geometric figures and alphabetic letters). In this paper, we will focus on the BOLD response for structured images in 1017 voxels belonging to primary visual cortex V1 only (see Fig. 1). Input and output data is standardized prior to applying sparse OPLS.

In our experiments, we examined how sparse OPLS results change as a function of the number of latent variables  $K$  as well as the regularization parameters  $\nu$  and  $\Lambda$ . The effect of  $\Lambda$  is examined for the diagonal case  $\Lambda = \lambda I_P$ , which amounts to a ridge penalty, and for the non-diagonal case  $\Lambda = \lambda R$ , which is used to induce spatial smoothing. The structure matrix  $R$  is specified as follows

$$R_{ij} = \begin{cases} 1 & \text{if } i = j \\ -\frac{1}{n_j} & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

where  $i \sim j$  holds if voxels  $i$  and  $j$  are neighbours and  $n_j$  represents the number of neighbours of  $j$ . Neighbours of a voxel  $i$  are taken to be those voxels which are adjacent to  $i$  in either the  $x$ ,  $y$  or  $z$  direction.

Reconstruction error is quantified as the average sum squared error (SSE) between real and reconstructed images, sparseness is quantified as the proportion of zero elements in  $B$  and spatial smoothness is quantified in terms of  $1/(\lambda \sum_k B_k^T R B_k)$ . All quantities were estimated using a ten-fold cross-validation scheme.

### 4.2 Results

#### 4.2.1 Reconstruction error

We start by evaluating how reconstruction error changes as a function of the number of latent variables  $K$  given default regularization parameters  $\nu = 0.01$  and  $\Lambda = \lambda I_P$  with  $\lambda = 0.01$ . Figure 2.A shows that the SSE decreases as the number of latent variables increases, although improvements are marginal for large  $K$ . In the following, we will use ten latent variables. The invariances which are encoded by these latent variables are depicted in Fig. 3.A. For this dataset, given the default regularization parameters, estimated components remain very similar to the principal components that are used to initialize  $A$ . Figure 3.B visualizes the components when more extreme values are chosen for the regularization parameters. As shown in Fig. 4, reconstructions based on the defaults parameter values are highly accurate.

Next, the influence of  $\nu$  on reconstruction error is examined. Figure 2.B shows that a minimum SSE is obtained around  $\nu = 0.001$ . However, since we are also interested in sparse solutions and SSE does not change much for small  $\nu$ , we use  $\nu = 0.01$ , which amounts to a sparseness of about 0.9. I.e., about one in ten voxels will enter the solution. Finally, we examine how reconstruction error changes as  $\lambda$  is varied for diagonal and non-diagonal  $\Lambda$ . Figure 2.C shows that SSE increases for large values of  $\lambda$ . Interestingly, the SSE is lower for non-diagonal than for diagonal  $\Lambda$ .

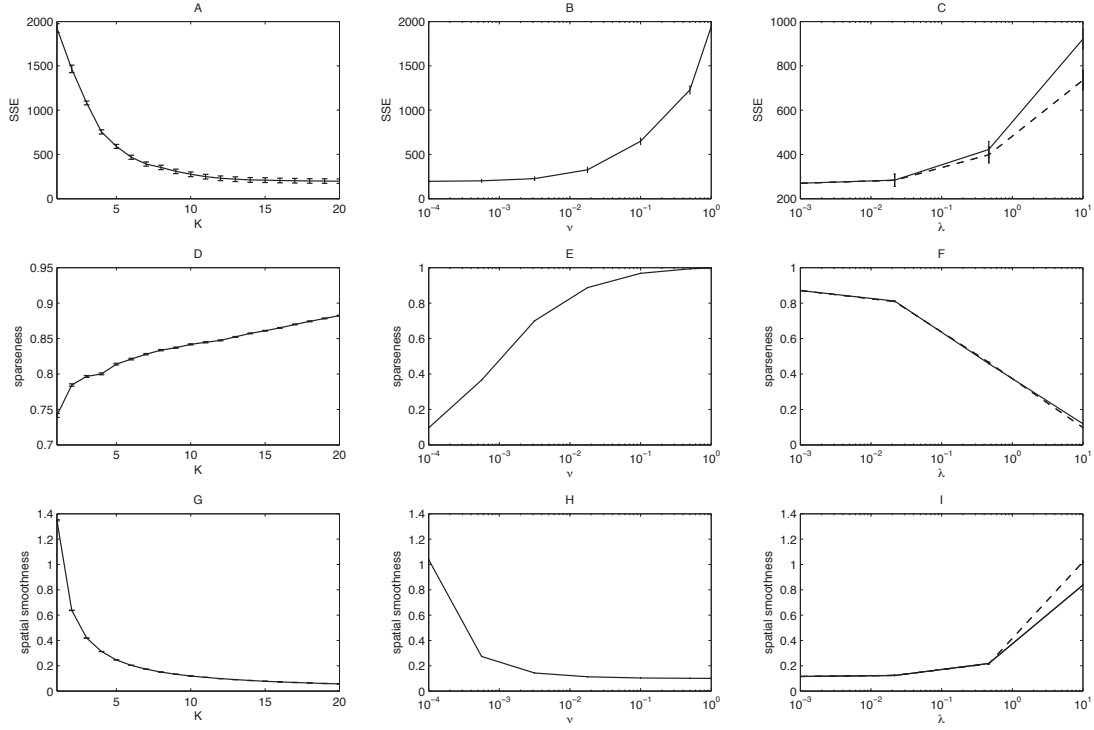


Figure 2: Reconstruction error (top row), sparseness (middle row) and spatial smoothness (bottom row) as a function of the number of latent variables  $K$  (left column), regularization parameter  $\nu$  (middle column) and regularization parameter  $\lambda$  (right column). Unmodulated parameters are fixed to their default values ( $K = 10$ ,  $\nu = 0.01$ ,  $\Lambda = 0.01I_p$ ). Solid lines correspond to  $\Lambda = \lambda I_p$  and dashed lines correspond to  $\Lambda = \lambda R$ . Results were computed using ten-fold cross-validation. Error bars denote the SEM.

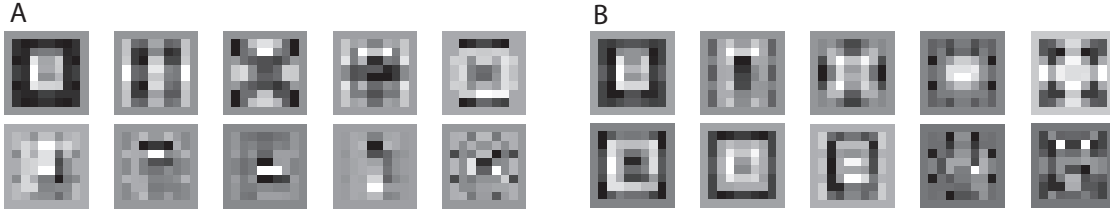


Figure 3: Invariances encoded by the first ten components of the loading matrix  $A$  given the defaults  $\nu = 0.01$  and  $\Lambda = 0.01I_p$  (panel A) and given extreme parameter values  $\nu = 1$  and  $\Lambda = R$  (panel B).

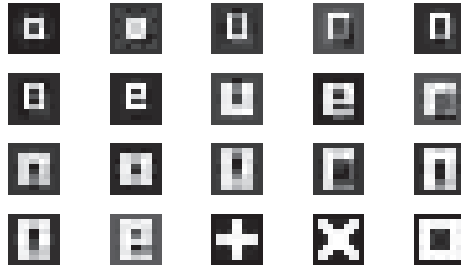


Figure 4: Twenty randomly selected images as reconstructed by sparse OPLS ( $K = 10$ ,  $\nu = 0.01$  and  $\Lambda = 0.01I_p$ ).

#### 4.2.2 Sparseness

We repeat the above analysis but now we evaluate sparseness. Figure 2.D shows that  $B$  becomes more and more sparse for increasing  $K$ . Sparseness also increases for large  $\nu$ , as expected (Fig. 2.E), but decreases for

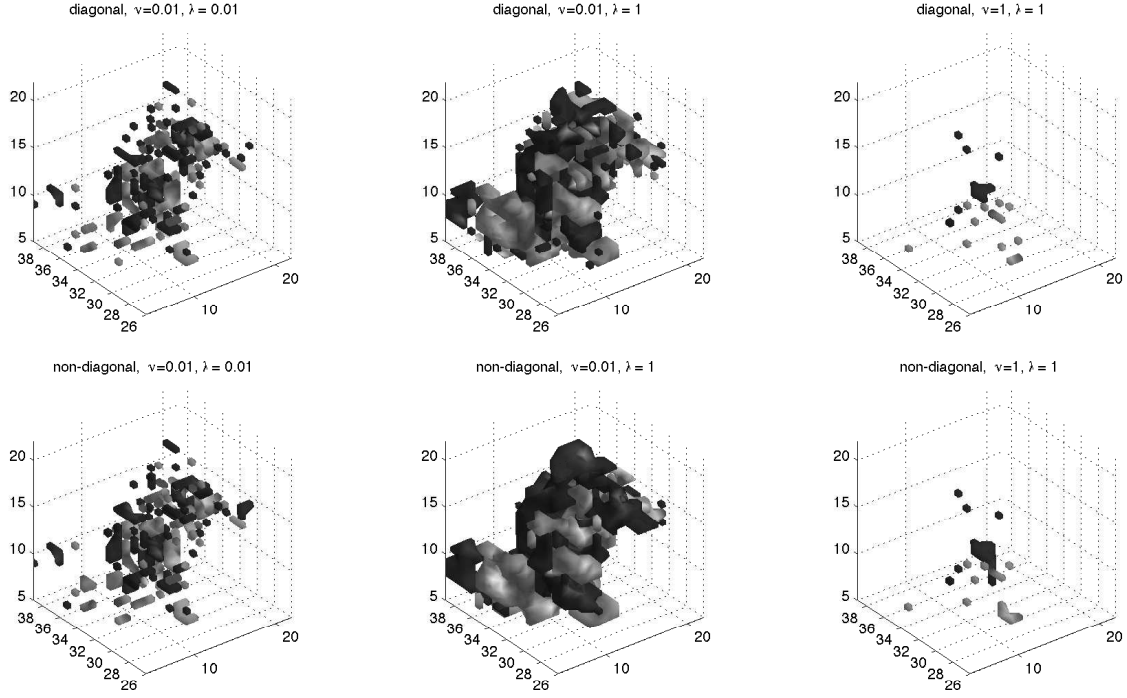


Figure 5: Selected voxels in  $B_1$  given different values of  $\nu$  and  $\lambda$  for diagonal and non-diagonal  $\Lambda$ . Positive coefficients are shown in light grey and negative coefficients are shown in dark grey. An increase in  $\lambda$  leads to the selection of more (correlated) voxels. Note that correlated voxels tend to lie close to each other, which already leads to a form of spatial smoothing. The non-diagonal  $\Lambda$  effectively creates connected blobs of negative and positive coefficients. An increase in  $\nu$  leads to sparser solutions.

large  $\lambda$  (Fig. 2.F). This is due to the fact that the elastic net tends to favor a large number of small coefficients instead of a small number of large coefficients in case of collinearity (Fig. 5). The diagonal  $\Lambda$  tends to give marginally sparser solutions than the non-diagonal  $\Lambda$  for large  $\lambda$ .

#### 4.2.3 Spatial smoothness

Spatial smoothness increases for individual  $B_k$  as  $k$  increases. This can be inferred from the decrease in smoothness in  $B$  which follows an exponential decay (Fig. 2.G). When  $\nu$  is increased, we observe a decrease in spatial smoothness (Fig. 2.H). This can be explained by the fact that voxels are gradually removed from the model which decreases smoothness since neighbouring voxels may still be part of the model. Spatial smoothness increases for increasing  $\lambda$ , both in the diagonal and the non-diagonal case (Fig. 2.I). This can again be explained by the properties of the elastic net since highly correlated neighbouring voxels will enter the model together (Fig. 5).

## 5 Discussion

In this paper, a sparse orthonormalized partial least squares algorithm is introduced. Sparseness in the loading matrix  $B$  is realized using a particularly efficient implementation of the elastic net algorithm while spatial smoothness is realized through the use of a non-diagonal regularizing matrix  $\Lambda$ .

The algorithm has been applied to the reconstruction of perceived images from BOLD response in primary visual cortex. Reconstructed images almost perfectly match their originals. In a sense, the reconstruction may have been too easy since the training data consisted of multiple measured responses to a small set of twenty structured images. As an alternative, we could have chosen to use random image data as input to our algorithm. However, we did not expect interesting components to be extracted from this data. Sparse OPLS is expected to work best in a regime where voxel subsets encode invariances in the output data, for example, in case of the reconstruction of perceived natural images [10].

## References

- [1] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.
- [2] H. Chun and S. Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal Of The Royal Statistical Society Series B*, 72(1):3–25, 2010.
- [3] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.
- [4] Y. Fujiwara, Y. Miyawaki, and Y. Kamitani. Estimating image bases for visual image reconstruction from human brain activity. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 576–584. 2009.
- [5] D. R. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. In *Sparsity and Inverse Problems in Statistical Theory and Econometrics Workshop*, 2008.
- [6] D. Hassabis, C. Chu, G. Rees, N. Weiskopf, P. D. Molyneux, and E. A. Maguire. Decoding neuronal ensembles in the human hippocampus. *Current Biology*, 19:546–554, 2009.
- [7] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:312–377, 1936.
- [8] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.
- [9] Y. Miyawaki, H. Uchida, O. Yamashita, M. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.
- [10] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.
- [11] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *SLSFS*, pages 34–51, 2005.
- [12] S. Roweis and C. Brody. Linear heteroencoders. Technical Report GCNU TR 1999-002, Gatsby Computational Neuroscience Unit, 1999.
- [13] C. Saunders, M. Grobelnik, S. R. Gunn, and J. Shawe-Taylor, editors. *Subspace, Latent Structure and Feature Selection, Statistical and Optimization, Perspectives Workshop*, volume 3940 of *Lecture Notes in Computer Science*, Bohinj, Slovenia, 2006. Springer.
- [14] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*, 67(1):91–108, 2005.
- [15] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.
- [16] C. Wang. Variational Bayesian approach to canonical correlation analysis. *IEEE Trans Neural Netw*, 18:905–910, 2007.
- [17] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67(2):301–320, 2005.
- [18] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.